An Experimentally Tested Scenario for the Structural Evolution of Eukaryotic Cys2His2 Zinc Fingers from Eubacterial Ros Homologs

Fortuna Netti, Gaetano Malgieri, Sabrina Esposito, Maddalena Palmieri, Ilaria Baglivo, Carla Isernia, James G Omichinski, Paolo V Pedone, Nicolas Lartillot, Roberto Fattorusso

An Experimentally Tested Scenario for the Structural Evolution of Eukaryotic Cys_2His_2 Zinc Fingers from Eubacterial Ros Homologs

Fortuna Netti,[1] Gaetano Malgieri,[1] Sabrina Esposito,[1] Maddalena Palmieri,[1] Ilaria Baglivo,[1] Carla Isernia,[1] James G. Omichinski,[2] Paolo V. Pedone,[1] Nicolas Lartillot,*[,2,3] and Roberto Fattorusso*[,1]

[1]Dipartimento di Scienze Ambientali, Biologiche e Farmaceutiche, Seconda Università di Napoli , Via Vivaldi 43, 81100 Caserta, Italy

[2]Département de Biochimie, Université de Montréal, Montreal, Québec, Canada

[3]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS-Université de Montpellier 2, Montpellier, France

***Corresponding authors:** E-mail: roberto.fattorusso@unina2.it; nicolas.lartillot@umontreal.ca.

Associate editor: Jeffrey Thorne

Abstract

The exact evolutionary origin of the zinc finger (ZF) domain is unknown, as it is still not clear from which organisms it was first derived. However, the unique features of the ZF domains have made it very easy for evolution to tinker with them in a number of different manners, including their combination, variation of their number by unequal crossing-over or tandem duplication and tuning of their affinity for specific DNA sequence motifs through point substitutions. Classical Cys_2His_2 ZF domains as structurally autonomous motifs arranged in multiple copies are known only in eukaryotes. Nonetheless, a single prokaryotic Cys_2His_2 ZF domain has been identified in the transcriptional regulator Ros from *Agrobacterium tumefaciens* and recently characterized. The present work focuses on the evolution of the classical ZF domains with the goal of trying to determine whether eukaryotic ZFs have evolved from the prokaryotic Ros-like proteins. Our results, based on computational and experimental data, indicate that a single insertion of three amino acids in the short loop that separates the β-sheet from the α-helix of the Ros protein is sufficient to induce a structural transition from a Ros like to an eukaryotic-ZF like structure. This observation provides evidence for a structurally plausible and parsimonious scenario of fold evolution, giving a structural basis to the hypothesis of a horizontal gene transfer (HGT) from bacteria to eukaryotes.

Key words: **zinc finger domain, nuclear magnetic resonance, phylogenetics analysis, evolution.**

Article

Fast Track

Introduction

Cys_2His_2-type zinc finger (ZF) domains are found in a wide variety of proteins that play important roles in regulating various cellular functions, including nucleic acid recognition, RNA packaging, apoptosis, protein assembly, and lipid binding (Laity et al. 2001). To date ZF domains represent the largest group of DNA-binding motifs known in eukaryotes (Klug and Schwabe 1995; Riechmann et al. 2000; Tupler et al. 2001; Esposito et al. 2006) and they are the founding member of a large family of zinc-binding modules. A key property of ZF domains is their autonomous folding pattern, which requires zinc coordination to stabilize the structure. The classical Cys_2His_2 ZF domain consists of approximately 25 amino acids arranged in a ββα topology in which a zinc ion is tetrahedrally coordinated by two cysteines and two histidines (Wolfe et al. 2000). The number of ZF domains present in proteins can vary between 1 and 37, but typically two or more ZF domains are required for high-affinity binding to a specific DNA target sequence. In certain cases, single ZF domains are sufficient for high-affinity binding to DNA, when they are associated with highly basic regions (Pedone et al. 1997; Omichinski et al. 1997; Dathan et al. 2002). The exact evolutionary origin of the ZF domain remains a mystery, as it is still not clear from which organisms it was first derived. However, the unique features of the ZF domains have made it very easy for evolution to tinker with them in a number of different manners, including their combination, variation of their number by unequal crossing-over or tandem duplication and tuning of their affinity for specific DNA sequence motifs through point substitutions. Through these evolutionary processes, it is conceivable that a large repertoire of transcription factors containing ZF domains could be created. Classical Cys_2His_2 ZF domains as structurally autonomous motifs arranged in multiple copies are known only in eukaryotes. Nonetheless, a single prokaryotic Cys_2His_2 ZF domain has been recently identified in the transcriptional regulator Ros from *Agrobacterium tumefaciens* (Chou et al. 1998, Malgieri et al. 2007). This single domain flanked by basic regions is able to bind DNA with high affinity and, despite utilizing a similar zinc coordination sphere, presents several important differences that distinguish it from the classical eukaryotic ZF. In particular, structural studies of the Ros ZF domain (Malgieri et al. 2007) have demonstrated that the ββα region that folds around the zinc ion is located within a larger globular domain of 58 amino acids, arranged in a

ββββαα topology, and stabilized by an extensive 15-residue hydrophobic core (Malgieri et al. 2007). A large number of Ros homologs have been found in different bacteria, mainly belonging to the α-proteo class. These proteins share high sequence identity with the Ros protein, but in several cases they either change the Cys_2His_2 zinc-coordination sphere or lose their zinc-binding property while still maintaining their ββββαα structure and DNA-binding activity (Baglivo et al. 2009; Russo et al. 2010). As far as the organismal macroevolutionary history is concerned, two different hypotheses have been proposed, both relying on horizontal gene transfer (HGT), although in opposite directions. The first (Bouhouche et al. 2000) postulates that eubacteria acquired Cys_2His_2 ZF domains by HGT from eukaryotes. This hypothesis was proposed at a time when Ros homologs had been identified only in a small number of bacterial species, most living in intimate ecological relationship with eukaryotes (e.g., symbiosis between an alpha-proteobacterium and a plant). Since then, Ros homologs have been identified in a large variety of bacterial species, most belonging to the alpha subdivision of the proteobacteria from which mitochondria originated, leading to the hypothesis that Cys_2His_2 ZF domains appeared as an innovation in proteobacteria, and were subsequently inherited via mitochondrial endosymbiosis (Moreira and Rodriguez-Valera 2000).

In this study, we provide computational and experimental evidence to elucidate the evolution of ZF domains assuming the hypothesis that the HGT occurred from prokaryotes to eukaryotes. In particular, we have performed phylogenetic, functional and structural analyses to test possible mutational events and structural modifications that could have been coupled to the HGT transfer. Our results indicate that a single insertion of three amino acids in the short loop that separates the end of the second strand of the β-sheet from the beginning of the α-helix of the Ros protein is sufficient to induce a structural transition from a Ros like to an eukaryotic -ZF like structure. This observation provides an experimental basis for a scenario of structural evolution from prokaryotic to eukaryotic ZFs occurring in very few steps, all involving structurally viable intermediates.

Results

The Frame-Shift Hypothesis

The Ros sequence was compared with a protein database containing multiple sequences via a BLAST search (Altschul et al. 1990). The resulting homologous sequences were aligned and used for phylogenetic analysis. This analysis has been conducted under both C20 mixture model (supplementary fig. S1A, Supplementary Material online) and LG empirical one-matrix model (supplementary fig. S1B, Supplementary Material online). Supplementary figure S1, Supplementary Material online, shows that members of this gene family are widely distributed among alpha-proteobacteria (168 proteins) and also delta-proteobacteria (59 proteins). Interestingly, two of the Ros homologs are proteins of eukaryotic origin possibly derived from recent HGT episodes (supplementary fig. S1, Supplementary Material online). In particular, one from *Ricinus communis* was correlated to *Caulobacter sp. K31* and the other from *Sordaria macrospora* was closely related to *Sphingomonadales*. The two models lead to similar outcomes, therefore, we choose the tree reconstructed under C20 model (supplementary fig. S1A, Supplementary Material online) because the small size of the alignment did not allow us to estimate the mixture directly on the data set.

The alignment of Ros homologs shows that several transitions in the zinc-coordinating patterns have occurred in the prokaryotic kingdom throughout evolution. In particular, there is a frequent occurrence of the CDHH coordination pattern as opposed to the more common CCHH coordination pattern. This observation suggests that an evolutionary mechanism introducing a cysteine point mutation three amino acids before the first zinc-coordinating residue could have transformed a bacterial ZF motif with the sequence $X_3CysX_2AspX_9HisX_{3\text{-}4}His$ into a $CysX_2CysX_{12}HisX_{3\text{-}4}His$ ZF motif (Baglivo et al. 2009) (here named amino acids frame-shift hypothesis). An aspartic acid residue in the second coordinating position may have favored such a transition because it possesses a lower affinity for zinc in comparison with cysteine (Baglivo et al. 2009).

As previously described (Baglivo et al. 2009), the CL protein (protein accession: ZP_03287372.1) contains atypical features that would appear to support the frame-shift evolutionary hypothesis of the ZFs. In addition to the residues that align with the zinc-coordinating residues in Ros87 (C_{81}, D_{84}, H_{94}, and H_{99}), the CL sequence contains an additional cysteine (C_{78}) located exactly three residues before the putative zinc-coordinating motif (supplementary fig. S2, Supplementary Material online). To evaluate whether the ZF domain of the CL protein could fold with a ββα topology similar to the eukaryotic ZF domains, a 33 amino acid peptide (named CL_CCDHH) encompassing the sequence of the ZF consensus of the CL protein (supplementary fig. S2, Supplementary Material online) was synthesized and structurally characterized. Upon addition of zinc ion (Isernia et al. 2003; Malgieri et al. 2011), the ^{1}H-NMR spectrum of CL_CCDHH displayed minimal chemical shift dispersion. This suggested that this peptide lacked stable tertiary structure and was unable to fold into the ββα topology. Moreover, a 2D-^{1}H ^{1}H NOESY of the peptide showed a limited number of signals in the NH–NH region of the spectrum and this is typical for what is observed for disordered peptides (supplementary fig. S3, Supplementary Material online). Again, this suggests that the peptide lacks a stable three-dimensional fold in the presence of zinc ion. The failure of the CL_CCDHH to fold into a domain structure even in the presence of zinc is may be due to the absence of a hydrophobic amino acid (typically F or Y) in the canonical positions (F/Y positions) of the eukaryotic ZF consensus sequence (fig. 1). In fact, we have analyzed more than 130 Cys_2His_2 ZF domain structures deposited in the PDB (Protein Data Bank) and verified that, in all cases, at least one of these two positions is occupied by an aromatic residue. In CL_CCDHH, these positions are occupied by an asparagine and a glycine, respectively, and it appears that these amino acids are not able to sufficiently stabilize the hydrophobic

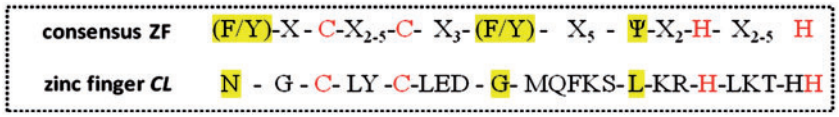


Fig. 1. Alignment of the primary sequence of the CL zinc-finger and the ZF consensus sequence: the putative zinc-coordination residues are in red, whereas the residues observed at the canonical aromatic positions of the consensus are yellow highlighted.

core required for forming the eukaryotic ββα fold. To obtain a better understanding of the CL tertiary fold, a CL mutant (CL_{58-149}) corresponding to Ros ZF DNA-binding domain (Ros87) was generated and characterized by NMR spectroscopy. The ^{1}H-^{15}N HSQC spectrum of CL_{58-149} displayed chemical shift dispersion in both the proton and nitrogen dimensions (supplementary fig. S4, Supplementary Material online), typical of a partially folded protein. Careful analysis of the spectrum ^{1}H-^{15}N HSQC J-18 (supplementary fig. S4, Supplementary Material online) indicated that chemical shifts for the side chains of only two histidines could be observed. In particular, only one histidine was found to be in the $N_{\delta 1}$-H tautomer form (i.e., with the $N_{\varepsilon 2}$ unprotonated) with both $N_{\varepsilon 2}$ and $N_{\delta 1}$ chemical shifts typical of a zinc-bound form. Finally, heteronuclear ^{1}H-^{15}N NOE (h-NOE) measurements revealed that only 30 residues in CL display some degree of rigidity in comparison with almost 60 residues in Ros87. Overall, the NMR results clearly indicate that CL is unable to adopt a tertiary fold similar to the one observed for the Ros87 protein. The lack of a defined structure for CL_{58-149} does not support the amino acid frame-shift hypothesis of ZF evolution.

The Insertion Hypothesis

To further analyze the evolution of ZF domains, we used the software package *HMMER* to align the HMM (Hidden Markov Model) prokaryotic profile obtained from the prokaryotic Ros homologs sequences against a eukaryotic ZF data set (fig. 2). The best-fitting (Viterbi algorithm) alignments of the eukaryotic sequences on the prokaryotic profile all involve a three amino acid insertion within the spacer region between the second and the third residues involved in zinc chelation (although not always at the exact same position), suggesting that the eukaryotic ZF domain was formed through such a three amino acids insertion (the insertion hypothesis). To more precisely position the three amino acids insertion within the spacer region, the empirical logo (Schneider and Stephens 1990) of the eukaryotic ZF domain was compared with the logo of the prokaryotic ZF domain (fig. 3). The phenylalanine residue (F_{31} in both the prokaryotic and eukaryotic logos) located four residues after the second zinc-coordinating residue (typically C) was highly conserved in both the eukaryotic and prokaryotic logos. Moreover, a leucine residue located three residues after the conserved phenylalanine in the prokaryotic logo (L_{34}) was conserved in the eukaryotic logo (L_{37}), but it is located six residues after the conserved phenylalanine. This comparison suggests that a simple three amino acids insertion between the conserved phenylalanine and leucine residues could have differentiate the prokaryotic and eukaryotic ZF domains.

To further define the potential site of a three amino acid insertion between the conserved phenylalanine and leucine residues, we compared both the primary and secondary structures of the prokaryotic and the eukaryotic logos. The prokaryotic secondary structure comparison was based on the βββααα secondary structure obtained for Ros87, and the eukaryotic secondary structure was the ββα fold found in numerous structures of eukaryotic ZF domains. This comparison revealed that both F_{31} and L_{34} were involved in the formation of secondary structures. The conserved phenylalanine is the last residue of the third β-strand in the βββααα prokaryotic ZF fold and the last residue of the second β-strand in the ββα eukaryotic ZF fold. In the case of the conserved leucine, it corresponds to the first position of the first α-helix in the βββααα prokaryotic ZF fold and to the third position of the α-helix in the ββα eukaryotic ZF fold. Together, these observations suggested three possible positions for the three amino acids insertion, FxxxKSLKRH, FKxxxSLKRH, and FKSxxxLKRH, where x represents any amino acid.

Characterization of the Ros87 Insertion Mutants

To assess the hypothesis that a three amino acids insertion occurred in the spacer region of the prokaryotic ZF domain, we designed three Ros87 mutants corresponding to the three possible insertions FxxxKSLKRH, FKxxxSLKRH, and FKSxxxLKRH. To choose the three amino acids to be placed at each of the insertion (x) positions, 220 eukaryotic ZF domains with the pattern XCX_9LKRH were aligned together. Owing to the large uncertainty about the topology and the root of the eukaryotic ZF phylogeny, we did not attempt to reconstruct the ancestral eukaryotic ZF sequence. Instead, we relied on a simple count-based method. Accordingly, the most frequent amino acid combinations in the resulting alignment for the three positions were KSR, SRG, and RGF for the first, the second and the third positions, respectively. Finally, a statistical analysis on the amino acid frequencies in the insertion positions between F_{31} and L_{34} of Ros87 suggested that the evolutionary insertion would have occurred three amino acids before L_{34} of the first α-helix (fig. 4).

The three Ros87 mutant proteins corresponding to the insertions selected above were purified, ^{15}N-labeled and analyzed by NMR spectroscopy. The ^{1}H-^{15}N HSQC spectra of all three mutants contained the expected number of signals as well as chemical shift dispersion patterns consistent with a stably folded domain (fig. 5 and supplementary fig. S5, Supplementary Material online). Preliminary analysis of the three spectra, based on acquisition of $^{1}H^{15}N$ HSQC spectra as a function on time, indicated that Ros87_RGF was the most stable mutant. To gain further structural insights, the H_N, H_α, C_α, and C_β chemical shifts assignments of the Ros87_RGF

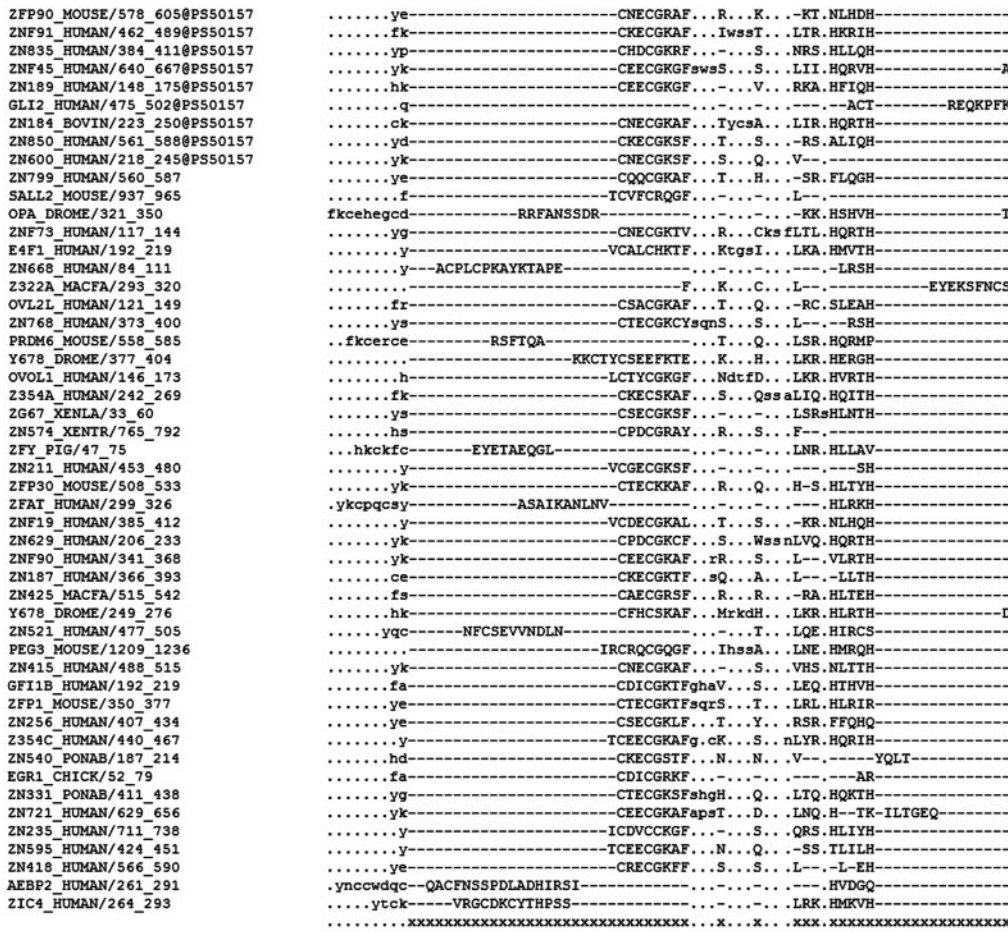


Fig. 2. Multiple alignment of 50 eukaryotic ZF domains and the prokaryotic ZF HMM profile. At the end of the block of Stockholm alignment, a line shows the reference coordinate annotation, with an x marking each column that the profile considered to be consensus; dots (.) in this line indicate insertions.

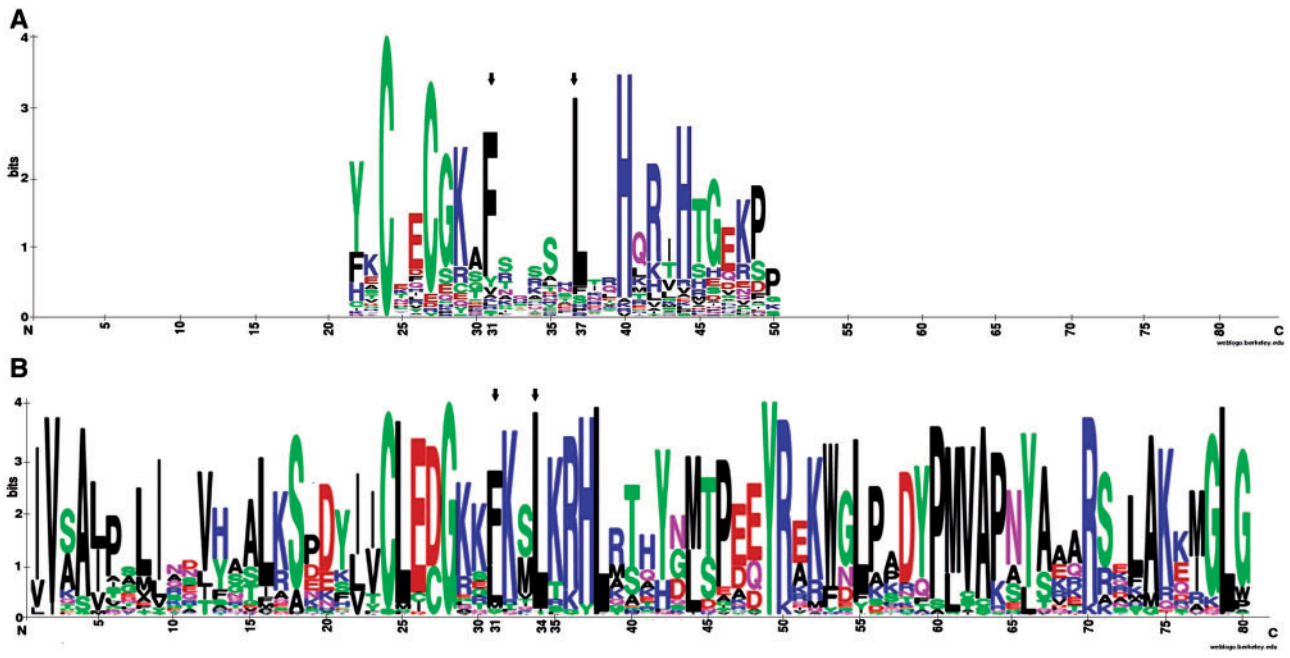
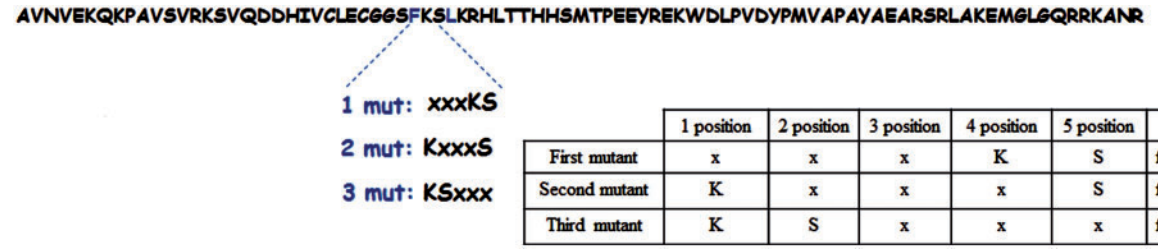


Fig. 3. Generated WebLogo (Suchard and Redelings 2006) for the eukaryotic ZF data set (*A*) and prokaryotic ZF data set (*B*). Positions 31 and 37 in (*A*) and 31 and 34 in (*B*) show highly conserved consensus residues. The logos are such that the height of each letter is proportional to the frequency of the corresponding amino acid and the overall height of each stack is proportional to the degree of sequence conservation at that position.



	1 position	2 position	3 position	4 position	5 position	% frequencies
First mutant	x	x	x	K	S	$f(K_4S_5)$= 0.08 %
Second mutant	K	x	x	x	S	$f(K_1S_5)$= 1.13 %
Third mutant	K	S	x	x	x	$f(K_1S_2)$= 2.83 %

FIG. 4. Primary sequence of Ros87 showing the amino acid insertion positions for the three Ros87 mutants. In the table, the amino acid frequencies in the mutant positions between F_{31} and L_{34} of Ros87 are showed.

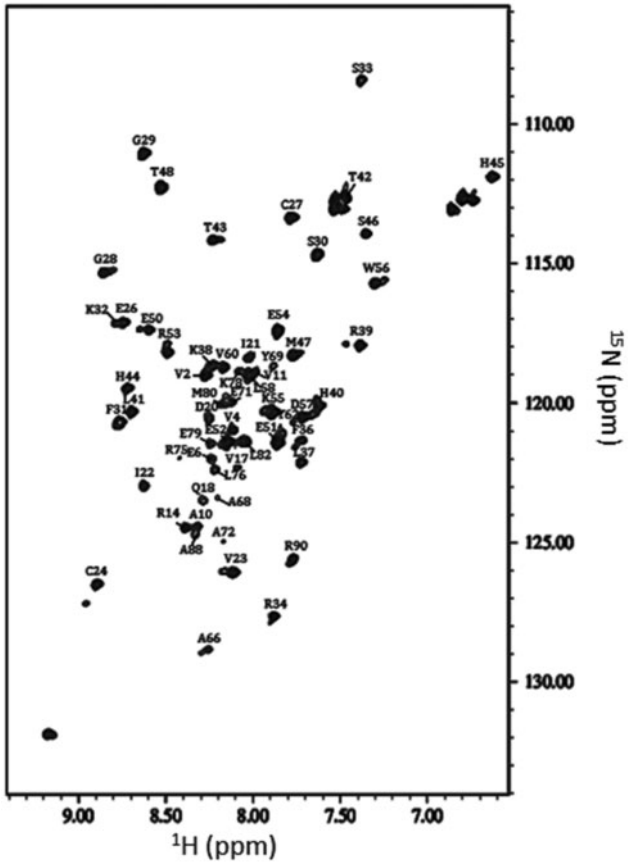


FIG. 5. ^1H-^{15}N HSQC spectrum of Ros87_RGF recorded at 298 K. Assignment of each unambiguously identified backbone resonance is reported.

mutant were obtained using a series of NMR experiments. The Chemical Shift Index (CSI) (Wishart et al. 1992) derived from the assignments allowed us to make a reliable prediction of the secondary structure elements present in the Ros87_RGF mutant (fig. 6). Interestingly, the protein contains a eukaryotic type ZF domain consisting of two β-strands (β1 = H_{21}–V_{23}; β2 = G_{29}–F_{31}) and an α-helix (R_{34}–H_{44}) with a two-residue loop between the second β-strand and the α-helix. In addition, there is a second 8-residue α-helix (α2 = P_{49}–D_{57}) that begins five residues after the end of the first α-helix as found in Ros87. However, the first β-strand present in the prokaryotic ZF domain is missing. Although the Ros87_RGF mutant maintains a significant portion of the original Ros87 ββααα structure, it appears to have lost the extensive 15-residue hydrophobic core. Furthermore, a similar secondary structure may be present in the other two insertion mutants, as their ^1H-^{15}N HSQC spectra display similar chemical shift patterns to the Ros87_RGF mutant protein (supplementary fig. S5, Supplementary Material online).

Next, we characterized the backbone dynamics of Ros87_RGF mutant and compared it with that of Ros87 protein. Based on heteronuclear ^1H-^{15}N NOE measurements (fig. 6), it appears that the backbone of the Ros87_RGF mutant protein is rigid over 43 residues between H_{21} and M_{64} (corresponding to M_{61} in Ros87) with an average value of 0.86 ± 0.04. In contrast, the first 20 residues at the N-terminus and the last 23 residues at the C-terminus display negative h-NOE values characteristic to highly mobile regions of the protein. These data indicate that the Ros87_RGF mutant protein has more flexible regions in comparison with the wild-type Ros, which is rigid over 58 residues (Malgieri et al. 2007). In addition, $^1H/^2H$ exchange experiments demonstrate that most of the amide protons in the Ros87_RGF mutant protein exchange within the first 20 min at pH = 6.8. In fact, only five residues (C_{24}, E_{26}, C_{27}, L_{41}, and side chain of W_{56}) displayed intermediate exchange rates with a time constant around 24 h. Under the same conditions, the Ros87 protein contains 38 residues that display intermediate exchange rates with time constants around 24 h. These differences in $^1H/^2H$ solvent exchange indicate diverse degrees of solvent exposure, and thus of backbone flexibility, further underlining the structural differences between the Ros87_RGF protein and the Ros87 protein.

These results clearly indicate that a three amino acids insertion in the spacer between the second and the third zinc-coordinating residues in Ros87 stabilizes a tertiary fold that possesses many characteristics of a canonical eukaryotic ββα motif with an additional helix at the C-terminus (fig. 7). Interestingly, the mutants no longer possess the ability to bind the cognate Ros DNA-binding sequence, which is consistent with the fact that most single eukaryotic ZF domains are not able to bind to DNA with high affinity and specificity unless accompanied by flanking basic regions in precise locations (supplementary fig. S7, Supplementary Material online).

Discussion

Protein domains posses stable three-dimension structures that are often more conserved than their sequences. However, progressive changes can occur during evolution, resulting in variations in both the sequence and the structure

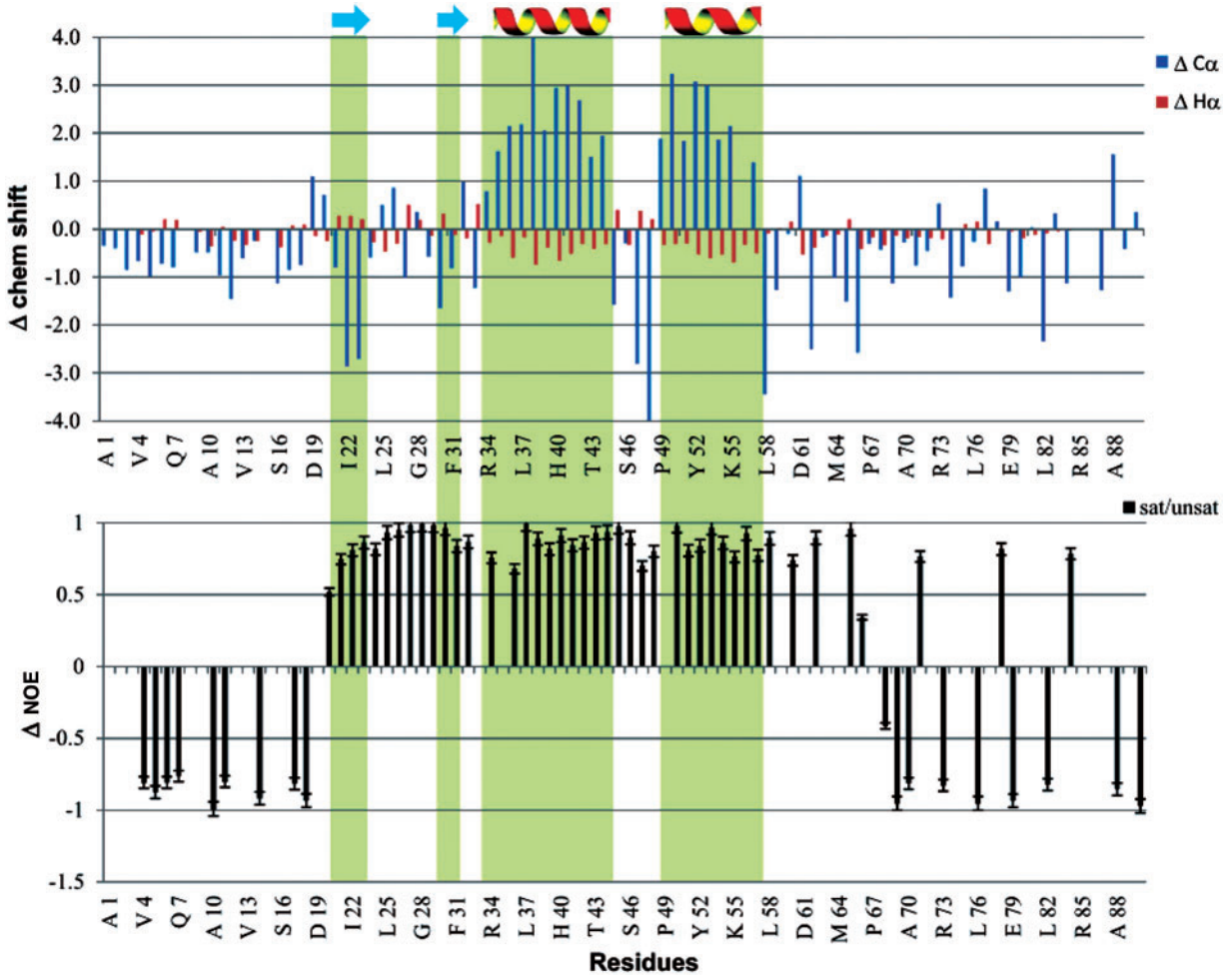


FIG. 6. (Top) Bar plot of Cα and Hα chemical-shift deviations from their random-coil values (δ observed $-$ δ random coil) as a function of residue number for Ros87_RGF protein. Secondary structure fragments are indicated. (Bottom) Bar plot of heteronuclear NOE values of Ros87_RGF protein. h-NOE values for some residues were not presented because of resonance overlap.

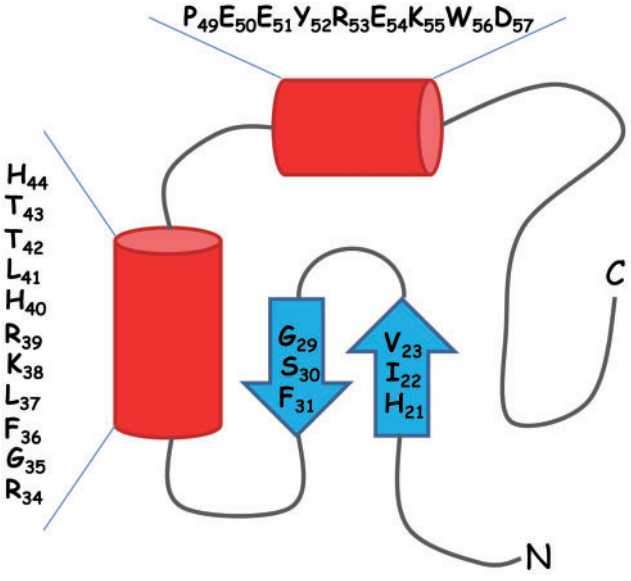


FIG. 7. Topology diagram for the structure of Ros87_RGF protein showing secondary structure motifs (α-helices and β-strands are depicted as red barrels and blue arrows, respectively). It is based on the NMR structure of Ros87 and the CSI of Ros87_RGF.

within a superfamily (Scheeff and Bourne 2005). Insertions, deletions, and single amino acid substitutions are the most common events in protein evolution and these events lead to the genesis of new protein folds (Grishin 2001). Our current work focused on the evolution of the classical ZF domains with the goal of trying to determine whether or not eukaryotic ZFs have evolved from the prokaryotic Ros-like proteins. To probe this possibility, we examined mutational events that could lead to an increase in the number of residues separating the second and third zinc-coordinating residues in the Ros protein from 9 to 12. More specifically, we have analyzed two possible evolutionary events that could result in the conversion, either a shift by three amino acids in the N-terminal direction to a new pair of zinc chelating residues (the frameshift hypothesis), or the insertion of three amino acids in the linker between the second and the third zinc-coordinating amino acids (the insertion hypothesis). In the CL protein (Chou et al. 1998), there is a cysteine residue (C_{78}) located exactly 3 positions before the first zinc-coordinating cysteine of CDHH Ros homologs. Zinc binding through this residue would result in a CCHH coordination pattern with 12 residues between the second and third chelating amino acids, as

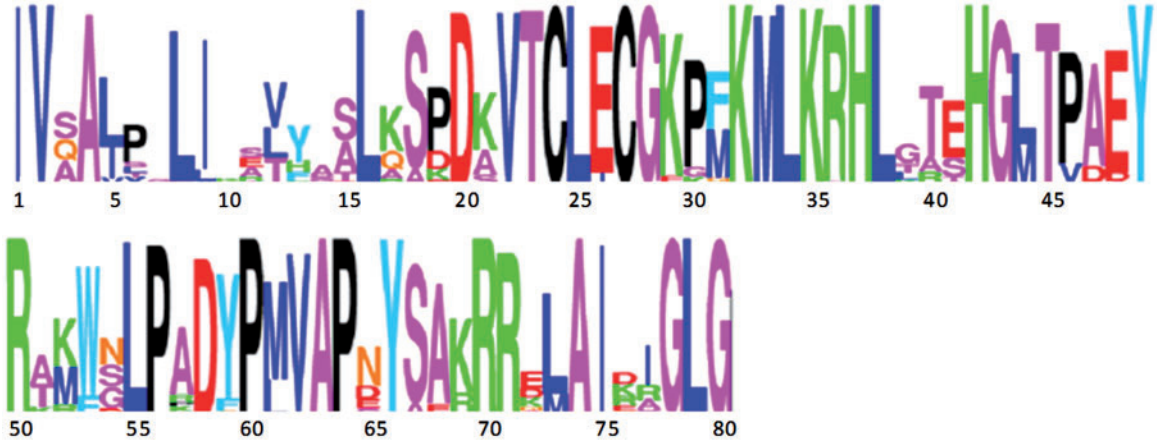


FIG. 8. Logo representation of the ancestor sequence of all *alpha proteobacteria*: the height of the letters is related to the percentage of amino acid variability present at that position.

proposed by the frameshift hypothesis. NMR characterization of the CL_CCDHH peptide showed that it does not possess the capacity to adopt a stable three-dimensional structure in the presence of zinc ions, which suggests that it is unable to stably coordinate the zinc ion in a tetrahedral geometry (supplementary fig. S3, Supplementary Material online). The inability of the CL protein to form a stably folded domain structure in the presence of zinc ions is most likely due to improper spacing of hydrophobic/aromatic amino acids required to form the core of the classical ZF domain structure (fig. 1). Moreover, ^{1}H-^{15}N HSQC spectra of the $CL_{58\text{-}149}$ protein indicate that the CDHH prokaryotic ZF domain present in CL is also not able to fold correctly (supplementary fig. S4, Supplementary Material online). This could be due to the presence of an extra cysteine, C_{78}, competing with C_{81} and D_{84} for the zinc coordination. Overall, our structural analysis of the CL protein clearly shows that at least two or more successive point mutations would be necessary to obtain a stably folded eukaryotic-like ZF motif within the CL sequence. Such a double mutational event, without a structurally viable intermediate, is a less parsimonious scenario of the evolution of the ZF domain from prokaryotes to eukaryotes.

Phylogenetic and sequence analyses indicate that the second hypothesis based on the three amino acids insertion, occurring as a single mutational event, is evolutionary and structurally more likely to have occurred. In fact, the comparison between the eukaryotic and prokaryotic logos (fig. 3) obtained using two data sets of eukaryotic and prokaryotic ZF domains demonstrates that the three amino acids insertion, responsible for the main difference between the two domains, might have occurred between the conserved residues F_{31} and L_{34} of a Ros-like CCHH protein. We have tested this insertion hypothesis and performed NMR spectroscopy studies on three Ros87 mutants (Ros87_KSR, Ros87_SRG, and Ros87_RGF) containing an insertion of three amino acids between these conserved residues. Interestingly, the NMR studies clearly indicate that the three insertions induce significant conformational changes. Importantly, however, these single insertion events were sufficient for the protein to adopt a unique tertiary fold that resembles to a large extent the eukaryotic ZF motif. Our experiments have several important evolutionary implications. First, they provide evidence for a structurally plausible and parsimonious scenario of fold evolution, giving a structural basis to the hypothesis of a HGT from bacteria to eukaryotes as postulated by Moreira and Rodriguez-Valera (2000). In certain respects, the fact that it is so easy to evolve from a prokaryotic Ros-like protein to a eukaryotic like ZF would even suggest the possibility that eukaryotic ZFs could have evolved multiple times from Ros homologs. Phylogenetic analysis of eukaryotic and bacterial ZFs (supplementary fig. S6, Supplementary Material online) suggests monophyly of eukaryotic ZFs, thus being more compatible with the idea of a unique origin of these ZFs. However, owing to the very small number of aligned positions, this phylogenetic tree is poorly supported, and does not allow us to rule out the alternative hypothesis in a definitive manner. Assuming a unique origin for eukaryotic ZFs, according to the structural scenario proposed here, one remaining question is whether this origin can be traced back to the mitochondrial endosymbiosis. Mitochondria are of alpha-proteobacterial origin, and most of the alpha-proteobacterial Ros homologs play important roles in developmental cycles (Gray et al. 1999). Interestingly, Bayesian ancestral sequence reconstruction (fig. 8) indicates that alpha-proteobacterial ZF proteins had ancestrally the canonical CCHH coordination pattern that we hypothesize was present in the original protein in which the three amino acid insertions may have occurred.

In conclusion, we have shown that a structural evolution from prokaryotic to eukaryotic ZFs may occur in very few steps: an insertion of three amino acids followed by a loss of the surrounding globular domain, and then tandem duplications and diversification of the domains.

Materials and Methods

Reagents

Apo-CCDHH peptide was purchased from *InBIOS*. CL_CCDHH-Zn(II) complex was obtained through the folding

procedure previously described (Isernia et al. 2003; Malgieri et al. 2011).

Phylogenetic Analysis

A blast search (Altschul et al. 1990) was performed to compare the Ros sequence (as query) with protein databases containing multiple such sequences (all nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF databases). Default parameters were used for the analysis, except for the *E-value* cutoff (value of the expected threshold) that was set as $\leq 10^{-6}$. The resulting sequences (419 sequences) were edited using Seaview (Gouy et al. 2010); the sites corresponding to the ZF domains were selected, and the alignment was manually adjusted. The alignment was filtered from poorly aligned positions using Gblocks (Castresana 2000; Talavera and Castresana 2007). The output file (containing 229 sequences), with conserved regions lacking the deleted positions in the sequences (blocks), was used for the following phylogenetic analysis. Phylogenetic trees were estimated using PhyloBayes 3.2 (Lartillot et al. 2009), under the C20 mixture model (Quang et al. 2008), and the LG empirical matrix model (Le et al. 2012). Two independent runs were performed under each model. Convergence was checked with the bpcomp program, the first 100 points were discarded as the burn-in, and the posterior consensus tree was computed over the remaining 19,842 trees (selecting one every two trees over all chains). Phylogenetic analyses of eukaryotic and prokaryotic sequences were performed using Bali-Phy 2.0.2 (Suchard and Redelings 2006) because this program has implementations to handle difficult-to-align sequences. A set of 20 prokaryotic Ros homologs sequences, randomly chosen from our data set by a home-made Perl script, has been added to a small representative set of 20 Cys_2His_2 eukaryotic sequences, chosen with the same procedure. This set has been used as input to calculate the phylogenetic tree, using the LG model.

Sequence Logo of ZFs

Sequence logos were obtained using WebLogo software (version 2.8.2, http://weblogo.berkeley.edu/) (Crooks et al. 2004). All sequences of prokaryotic Ros homologs (227 sequences) and a small representative set of 100 Cys_2His_2 eukaryotic ZF from PROSITE were used to generate the prokaryotic and eukaryotic logos, respectively.

Ros87 Mutants Design

The software package *HMMER* (Finn et al. 2011) was used to design the mutants. First, the available sequences of classical eukaryotic ZF proteins from the protein database PROSITE (12,405 sequences) were gathered, and a small representative set of 50 Cys_2His_2 eukaryotic ZF was selected at random using a home-made Perl script. Then, a profile HMM of the prokaryotic ZF domains was constructed using the alignment of the Ros homologs (74 sequences) as input to HMM build. The prokaryotic HMM profile was then aligned with the eukaryotic set. To determine where the amino acids would be inserted in the Ros87 mutants, the eukaryotic ZF proteins from PROSITE with the pattern XC_9LKRH (220 proteins) were selected using a Perl script. Another Perl script was used to count the occurrences of all amino acid triplets in the putative insertion positions.

Cloning and Purification of the Proteins

The coding sequence for the $CL_{58\text{-}149}$ protein ZP_03287372.1 was generated by polymerase chain reaction (PCR) from DNA extract by main vein of leaves of a citrus plant infected by *Candidatus Liberibacter asiaticus*, kindly provided by Dr Diva Teixeira (Fundecitrus, Araraquara, CEP, Brazil). The primers utilized were as follows: CandN58NdeIfor, 5′-GGAATTCCATATGaacgtgcaacctgaaaggctcaaacc-3′, and CandV149Stop SalIrev, 5′-aacatgGTCGACTCATACCTTAGAAGTCAATACACG-3′. PCR product was digested with the restriction enzymes NdeI and SalI and cloned into pET-22b(+) bacterial expression vector. DNA fragments encoding the three different Ros87 insertion mutants (Ros87_F31_K32insKSR and Ros87_K32_S33insSRG, Ros87_S33_L34insRGF) were generated by PCR using as template the plasmid carrying the coding sequence for Ros87 ($Ros_{56\text{-}142}$). All the mutants were generated by PCR-mediated mutagenesis according to the method of Delidow et al. (1993) using a double-step reaction. The following oligonucleotides were used as primers: Rosdel56NcoI, 5′-ACATGCCATGGCGGTCAATGTTGAAAAGCA-3′, Ros_F31insKSRfor, 5′-gtttggaatgtggtggctcgttcaagtcgcgcaagtcgctcaaacgccacctg-3′, Ros_F31insKSRrev, 5′-CAGGTGGCGTTTGAGCGACTTGCGCGACTTGAACGAGCCACCACATTCCAAAC-3′, and RosR142, 5′-CGGAATTCTCAACGGTTCGCCTTGCGG-3′, for Ros87_KSR; Rosdel56NcoI, RosRos_K32insSRGfor, 5′-gaatgtggtggctcgttcaagtcgcgcggctcgctcaaacgccacctgacg-3′, Ros_K32insSRGrev, 5′-CGTCAGGTGGCGTTTGAGCGAGCCGCGCGACTTGAACGAGCCACCACATTCC-3′, and RosR142 for Ros87_SRG; Rosdel56NcoI, Ros_S33insRGFfor, 5′-gtggtggctcgttcaagtcgcgcggcttcctcaaacgccacctgacgacgc-3′, Ros_S33insRGFrev, 5′-GCGTCGTCAGGTGGCGTTTGAGGAAGCCGCGCGACTTGAACGAGCCACCAC-3′, and RosR142 for Ros87_RGF. The obtained DNA fragments were then digested with the restriction enzymes NcoI and BamHI, cloned into the pET-11d vector. ^{15}N- and ^{15}N-^{13}C-labeled proteins were overexpressed and purified as previously published (Esposito et al. 2006) with the only exception that the protein expression for Ros87_SRG and Ros87_RGF was induced for 1.5 h at 22 °C.

NMR Spectroscopy

The 1D NMR spectrum of the CL_CCDHH peptide was acquired with 16 K data points and was zero filled before the Fourier transform. 2D TOCSY (Braunschweiler and Ernst 1983) and 2D NOESY experiments (Jeener et al. 1979; Wuthrich 1986) were acquired with a mixing time of 70 ms and 100 and 150 ms, respectively. NMR samples of unlabeled or labeled $CL_{58\text{-}149}$ and Ros87 mutants typically contained 1 mM of the protein, 20 mM phosphate buffer, 4 mM TCEP (only in $CL_{58\text{-}149}$ samples), 0.2 M NaCl dissolved in $H_2O/^2H_2O$ 90:10 at pH 6.8. A 100% ^2H_2O was used for the $^1H/^2H$ exchange experiments. $^1H^{15}N$ HSQC J-18 spectra were acquired

as previously described (Esposito et al. 2006). To obtain sequence-specific backbone and C_β resonances assignment, a standard set of triple-resonance NMR experiments was collected on a ^{15}N-^{13}C labeled sample of Ros87_RGF ins.

The heteronuclear NOE effect was measured with standard refocused HSQC pulse sequence in the presence or absence of proton decoupling during the 5-s relaxation delay on a ^{15}N-Ros87_RGF sample. Hetero-NOE values were derived from the intensity ratios of the cross-peak with and without proton decoupling. The spectra analysis was performed using VNMR1 6.1B (Varian) and XEASY (Bartels et al. 1995) programs.

Gel Mobility Shift Analysis

Five pmol of the purified proteins were incubated for 10 min on ice with 2 pmol of the duplex oligonucleotide VirC 5′-GATTTTATATTTCAATTTTATTGTAATATAATTTCAATTG-3′, in the presence of 25 mM HEPES (pH 7.9), 50 mM KCl, 6.25 mM $MgCl_2$, 1% NP-40, and 5% glycerol. After incubation, the mixture was loaded on a 5% polyacrylamide gel (29:1 acrylamide:bisacrylamide ratio) and run in 0.5% TBE (200 V for 75 min). The gels were than stained with SYBR Green and visualized with the Typhoon Trio $^{++}$ scanner (GE Healthcare).

Supplementary Material

Supplementary figures S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr Vincenzo Piscopo, Mr Maurizio Muselli, and Mr Marco Mammucari for the excellent technical assistance. Computational resources were provided by Calcul Québec, Compute Canada, and the Canadian Foundation for Innovation. This work was partially supported by the Ministero dell'Istruzione, dell'Università e della ricerca (MIUR) Programma MERIT RBNE08HWLZ_014, PRIN 2010 2010M2JARJ_002, and the Natural Science and Engineering Research Council of Canada to N.L. This work also was supported in part by a grant from the National Science and Engineering Research Council of Canada to J.G.O.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Baglivo I, Russo L, Esposito S, Malgieri G, Renda M, Salluzzo A, Di Blasio B, Isernia C, Fattorusso R, Pedone PV. 2009. The structural role of the zinc ion can be dispensable in prokaryotic zinc-finger domains. *Proc Natl Acad Sci U S A.* 106:6933–6938.

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K. 1995. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biol NMR.* 6:1–10.

Bouhouche N, Syvanen M, Kado CI. 2000. The origin of prokaryotic C2H2 zinc finger regulators. *Trends Microbiol.* 8(2):77–81.

Braunschweiler L, Ernst RR. 1983. Coherence transfer by isotropic mixing: application to proton correlation spectroscopy. *J Magn Reson.* 53(3):521–528.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17(4):540–552.

Chou AY, Archdeacon J, Kado CI. 1998. Agrobacterium transcriptional regulator Ros is a prokaryotic zinc finger protein that regulates the plant oncogene ipt. *Proc Natl Acad Sci U S A.* 95(9):5293–5298.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14(6):1188–1190.

Dathan N, Zaccaro L, Esposito S, Isernia C, Omichinski JG, Riccio A, Pedone C, Di Blasio B, Fattorusso R, Pedone PV. 2002. The *Arabidopsis* SUPERMAN protein is able to specifically bind DNA through its single Cys2-His2 zinc finger motif. *Nucleic Acids Res.* 30(22):4945–4951.

Delidow BC, Lynch JP, Peluso JJ, White BA. 1993. Polymerase chain reaction: basic protocols. *Methods Mol Biol.* 15:1–29.

Esposito S, Baglivo I, Malgieri G, et al. (11 co-authors). 2006. A novel type of zinc finger DNA binding domain in the *Agrobacterium tumefaciens* transcriptional regulator Ros. *Biochemistry* 45(34):10394–10405.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39(Web Server issue):W29–W37.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. *Science* 283(5407):1476–1481.

Grishin NV. 2001. Fold change in evolution of protein structures. *J Struct Biol.* 134(2–3):167–185.

Isernia C, Bucci E, Leone M, et al. (12 co-authors). 2003. NMR structure of the single QALGGH zinc finger domain from the *Arabidopsis thaliana* SUPERMAN protein. *Chembiochem* 4(2–3):171–180.

Jeener J, Meier BH, Bachmann P, Ernst RR. 1979. Investigation of exchange processes by two-dimensional NMR spectroscopy. *J Chem Phys.* 71(11):4546–4553.

Klug A, Schwabe JW. 1995. Protein motifs 5. Zinc fingers. *FASEB J.* 9(8):597–604.

Laity JH, Lee BM, Wright PE. 2001. Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol.* 11(1):39–46.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286–2288.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol.* 29(10):2921–2936.

Malgieri G, Russo L, Esposito S, Baglivo I, Zaccaro L, Pedone EM, Di Blasio B, Isernia C, Pedone PV, Fattorusso R. 2007. The prokaryotic Cys2His2 zinc-finger adopts a novel fold as revealed by the NMR structure of *Agrobacterium tumefaciens* Ros DNA-binding domain. *Proc Natl Acad Sci U S A.* 104(44):17341–17346.

Malgieri G, Zaccaro L, Leone M, et al. (12 co-authors). 2011. Zinc to cadmium replacement in the *A. thaliana* SUPERMAN Cys(2) His(2) zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers* 95(11):801–810.

Moreira D, Rodriguez-Valera F. 2000. A mitochondrial origin for eukaryotic C2H2 zinc finger regulators? *Trends Microbiol.* 8(10):448–450.

Omichinski JG, Pedone PV, Felsenfeld G, Gronenborn AM, Clore GM. 1997. The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nat Struct Biol.* 4(2):122–132.

Pedone PV, Omichinski JG, Nony P, Trainor C, Gronenborn AM, Clore GM, Felsenfeld G. 1997. The N-terminal fingers of chicken GATA-2 and GATA-3 are independent sequence-specific DNA binding domains. *EMBO J.* 16(10):2874–2882.

Quang le S, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317–2323.

Riechmann JL, Heard J, Martin G, et al. (17 co-authors). 2000. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290(5499):2105–2110.

Russo L, Palmieri M, Baglivo I, Esposito S, Isernia C, Malgieri G, Pedone PV, Fattorusso R. 2010. NMR assignments of the DNA binding domain of Ml4 protein from *Mesorhizobium loti*. *Biomol NMR Assign.* 4(1):55–57.

Scheeff ED, Bourne PE. 2005. Structural evolution of the protein kinase-like superfamily. *PLoS Comput Biol.* 1(5):e49.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18(20):6097–6100.

Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* 22(16): 2047–2048.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56(4):564–577.

Tupler R, Perini G, Green MR. 2001. Expressing the human genome. *Nature* 409(6822):832–833.

Wishart DS, Sykes BD, Richards FM. 1992. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* 31(6):1647–1651.

Wolfe SA, Nekludova L, Pabo CO. 2000. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct.* 29: 183–212.

Wuthrich K. 1986. NMR of proteins and nucleic acids. New York: Wiley.